3.0 Data Selection Process

This section describes the data analysis process used to obtain concentration inputs to the screening assessment models from the raw concentration data. This process was repeated for each segment and for each contaminant being evaluated. The process involved choosing a maximum representative value for the concentration of each contaminant for a deterministic run and calculating the parameters that define the concentration probability density function needed for the stochastic runs. The term "maximum representative value" is used to mean the highest concentration value that is considered representative of the sampling location.

Because of the volume of data, it was desirable to keep human intervention in the data selection process to a minimum. PNNL guided the development of computational techniques to select the data values to be used in the screening assessment. Graphical displays of the raw data were used to track the results of the computational techniques and identify needed modifications to final data files. The data plots are included as Appendix C in Volume II. For a copy of Volume II (over 500 pages plus 9 diskettes), contact S. D. Cannon at 509-372-6210.

The work of the data task did not include analyzing the quality of the data. Data quality objectives will be discussed in the report on the screening assessment and requirements for a comprehensive assessment.

3.1 Deterministic and Stochastic Analyses

To meet the needs of the screening assessment, data were prepared to support two types of analyses:

- A **deterministic analysis** wherein a single calculation is performed with a single value selected for each parameter, such as the concentrations of contaminants entering the river.

- A **stochastic analysis** wherein a set of calculations is performed over the range of some of the input parameters.

For the deterministic analysis, the maximum representative concentrations for each contaminant in each medium were used to represent the segment. For the stochastic analysis, a probability density function of the concentration parameter will be assumed to be a lognormal distribution truncated at the 99.9th percentile. The probability density function expresses the state of knowledge about alternative values for the parameter. The particular lognormal distribution assumed will be determined by specifying the geometric mean and geometric standard deviation that represents concentration data for each river segment for which impacts will be computed.

3.2 Data Evaluation Conventions

Some general data evaluation conventions were applied in the course of selecting the data and preparing it for the automated processing. These conventions are the result of CRCIA Team decisions and EPA risk assessment guidance (EPA 1990).

- Both filtered and unfiltered data were used in the data selection process.

- If any datum was reported by the laboratory as "less than" the equipment detection limit, then the data value (which is the equipment detection limit) was replaced with half the reporting limit for that datum.

- Any data value labeled with a laboratory qualifier of "R" for "rejected" was removed from the data set. These data values were marked for rejection by the laboratories for reasons such as exceeding the holding times.

- If no detection limit information was provided, any data value labeled with a laboratory qualifier of "U" was removed from the data set. Data values labeled "U" by the laboratory were below the reporting limit for the constituent.

- Data values labeled with a laboratory qualifier of "U" were not used when choosing maximum values.

3.3 Process Used to Select Groundwater Data

One of the key parameters in the screening assessment calculations is the concentration of the contaminants in the groundwater entering the Columbia River from the Hanford Site. For the purpose of the screening assessment, Hanford groundwater data have been compiled that represent water quality in the upper part of the unconfined aquifer. This hydrologic unit offers the most direct pathway for contaminants to reach humans and environmental receptors. It also is believed to contain the majority of contaminants. Where data exist for wells that monitor deeper zones, the general pattern is lower concentrations or non-existent contamination with increasing depth in the aquifer (e-mail message from R.E. Peterson, ERC, to T.B. Miley, PNNL, March 20, 1996).

3.3.1 Selection of Groundwater Wells

The first step in selecting the set of groundwater wells as sources of data for the screening assessment was to identify all wells that have been sampled over the time period of interest. The next step was to use the Geographic Information System to determine which of these wells fell within the groundwater corridors specified for study. The well drilling information was then examined to determine the screening depth of the wells. Groundwater data from wells that monitor zones deeper than the upper unconfined aquifer were not included in the data compilation. Table 3.1 gives the list of wells by segment that were eliminated because they monitor at depths below the upper unconfined aquifer.

Table 3.1. Groundwater Wells Eliminated as Sources of Data for the Screening Assessment

Segment	Well Name
2	199-B2-12
2	199-B3-2P
4	199-K-32B
6	199-N-69
6	199-N-80
6	199-N-8P
8	199-D8-54B
8	199-D8-54B
10	199-H4-12C
10	199-H4-15C
10	199-H4-2
13	199-F5-43B
13	199-F5-43B
20	399-1-16C
20	399-1-17C
20	399-1-9
21	699-S29-E16C

3.3.2 Selection of Data

This section presents the data selection process for groundwater data. The process was followed for each contaminant and river segment.

3.3.2.1 Process the Raw Data for Inconsistencies

All groundwater data used in the assessment were from HEIS. ERC supplied a Microsoft Access macro to process the data for inconsistencies. The macro performed a series of queries designed to implement data standardization and to correct known errors. Soil results that reside in the groundwater area of HEIS were removed. Units of measurement were standardized for various contaminants. Known errors in constituent identification codes were corrected. Constituent names were standardized. Known errors in some of the nitrate analyses were corrected.

3.3.2.2 Identify at Most One Outlier

For each well, Dixon's test (Barnett and Lewis 1994) was conducted to decide if the largest data value is an outlier. The test assumes that the distribution (probability density function) of the data is normal except possibly for the potential single outlier. Because the groundwater data are assumed to be lognormally distributed, the data were log-transformed before this test was applied. If the data are lognormal, the log-transformed data will be normal as required by the test. When the data values were zero or negative, they were replaced with a small positive value for the log transformation.

The Dixon test examines the ratio between the difference in the largest and second largest data values and the range of the data. If this ratio is large, then the largest point is declared to be an outlier. The confidence level used for the test was 0.05. Any data identified as an outlier by the Dixon test received individual attention to determine whether they should indeed be deleted from the data set. This was done through a review of the data plots in Appendix C of Volume II. For a copy of Volume II (over 500 pages plus 9 diskettes), contact S. D. Cannon at 509-372-6210.

3.3.2.3 Test for a Trend Over Time

After any outlier was removed, the concentration data were tested for an upward or downward trend over time using the Mann-Kendall test (Gilbert 1987). To determine what data value is representative of current conditions in a well, it was necessary to know if the well data were trending over time. The Mann-Kendall test can be used regardless of the underlying data distribution. To perform the test, the data were ordered by sample date, then the sign of the difference (plus or minus) between each measurement and all subsequent measurements were calculated. The Mann-Kendall statistic is the number of positive differences minus the number of negative differences. If the test statistic was a large positive number, then measurements taken later in time were larger than those taken earlier, and an upward trend was present. If the test statistic was a large negative number, then the measurements taken later in time were smaller than those taken earlier, and a downward trend was present. A significance level of 0.01 was used in testing for an upward or downward trend in the concentration data. That is, if in fact there was no underlying trend in the data, 1 percent of the time the test would incorrectly indicate that a trend did exist.

3.3.2.4 Choose Representative Well Data

To support the deterministic and stochastic analyses, two representative values were selected for each well. These were a representative maximum and a representative median. Representative well data are selected after any outlier is removed. For a well that had no trend in its data, no data value is considered more representative than any other data value, so the representative values were selected based on all of the data values. For a well with a trend in the data values, the most recent data were considered most representative of the current conditions in the well.

3.3.2.5 Choose a Representative Well Maximum

A representative maximum value is used for the deterministic analysis because the goal is to produce a conservative or worst case estimate of risk. For each groundwater well with non-trending data, the maximum concentration detected in the well is chosen for the representative maximum value. For upward trending data, the maximum concentration was more conservative than the maximum of the current time period, so the overall maximum was used. For downward trending data, the most recent detected measurement was used. If there is more than one detected measurement in the most recent sampling period (as in one filtered and one unfiltered), then the maximum of the measurements is chosen as the representative maximum value.

3.3.2.6 Choose a Representative Well Median

A representative median value was used in the calculation of stochastic parameters because the stochastic process requires that attention be focused on best-estimate parameter values rather than conservative (maximal) values. If an upward or downward trend was detected, the median of the most recent groundwater concentration measurements was used to represent the well. If a well does not have a trend, then no single data point is considered more representative of the well than any other point. In that case, the median of the data is the single most representative concentration value for the well. This approach leads to the most representative probability density function to describe the uncertainty about the concentration data for the river segment being studied.

3.3.2.7 Compute Segment Parameters

For the groundwater medium, the representative values for individual wells must be combined into parameters that are representative of the river segment because no single well is representative of the segment. Whereas the process for the wells combines the data over time into a single value at the various well locations, the segment process combines the values over space into representative data for the segment.

3.3.2.8 Compute the Segment Maximum

The segment maximum is the highest of the well representative maximum values. This value is the maximum of all the observed concentrations in any well in the segment. This value is used to represent the segment in the deterministic risk calculations.

3.3.2.9 Compute Stochastic Parameters

The stochastic parameters (the geometric mean and geometric standard deviation) were calculated from the set of median (best-estimate) well values in the segment. The first step in calculating the geometric mean and geometric standard deviation was to take the natural logarithm of the median well values. The arithmetic mean of the log-transformed median values was calculated and then exponentiated to obtain the geometric mean. To calculate the geometric standard deviation, the standard deviation of the log-transformed median values was calculated and then exponentiated.

When some of the data are less than or equal to zero, the winsorized mean and standard deviation are computed (Dixon and Tukey 1968). Winsorizing is used to estimate the mean and standard deviation of a symmetric distribution even though the data set has a few missing or unreliable values at either or both ends of the ordered data set. The unreliable data at the lower end of the data set are replaced with the next largest data value, and an equal number of data are replaced at the upper end of the data set with the next smallest data value.

3.4 Process Used to Select Data for Other Media

For the sediment, seeps, surface water, and external radiation media, sampling locations within a segment cannot be easily pinpointed. Sampling locations tend to be regions rather than distinct locations such as a well. Also, in any one sampling period, there are few sampling events within a segment. Because the sampling does not occur at discrete locations for multiple times, it is not necessary to combine the data for a sampling location over time before calculating segment values.

3.4.1 Process the Raw Data for Inconsistencies

The data for all media other than groundwater were processed to remove inconsistencies. Units were standardized for various contaminants, and constituent names were standardized.

3.4.2 Identify at Most One Outlier

For each segment, Dixon's test (Barnett and Lewis 1994) was conducted to decide if the largest datum was an outlier. This test was applied to the set of all data over all sampling locations in the segment. As with the groundwater data, the data were log-transformed before this test was applied. The Dixon test used was described in the groundwater processing section above. When the data values were zero of negative, they were replaced with a small positive value for the log transformation. Any data identified as an outlier by the Dixon test received individual attention to determine whether they should indeed be deleted from the data set. This was done through a review of the data plots in Appendix C of Volume II. For a copy of Volume II (over 500 pages plus 9 diskettes), contact S. D. Cannon at 509-372-6210.

3.4.3 Compute the Segment Maximum

After removing at most one outlier, the maximum detected concentration was selected as the segment maximum. This value will be used for the deterministic screening assessment calculations.

3.4.4 Compute Stochastic Parameters

Calculate the geometric mean and geometric standard deviation of all measurements for the segment after outliers for the segment have been removed. The calculation of the geometric mean and geometric standard deviation are as described in section 3.3.2.5.2 in the groundwater process using winsorized data. The geometric mean and geometric standard deviation define the specific two-parameter lognormal distribution that will be used for the stochastic risk assessment calculations for the segment.